A New High Resolution Depth Map Estimation System Using Stereo Vision and Depth Sensing Device

Shuai Zhang, Chong Wang and S. C. Chan
Department of Electrical and Electronic Engineering,
The University of Hong Kong
{szjeff, cwang, scchan}@eee.hku.hk

***Abstract*—Depth map estimation is a classical problem in computer vision. Conventional depth estimation relies on stereo/multi-view matching or depth sensing devices alone. In this paper, we propose a system which addresses high resolution and high quality depth estimation based on joint fusion of stereo and Kinect data. The problem is formulated as a maximum a posteriori probability (MAP) estimation problem and reliability of two devices are derived. The depth map estimated is further refined by color image guided depth matting and a 2D polynomial regression (LPR)-based filtering. Experimental results show that our system can provide high quality and resolution depth map, which complements the strengths of stereo vision and Kinect depth sensor.**

***Index Terms*— Depth estimation system, high resolution, Kinect, stereo vision.**

I. INTRODUCTION

Depth information is an important ingredient in many advanced video applications such as image-based rendering (IBR) [3], object reconstruction, human computer interface, etc. Traditional depth acquisition is mainly based on either stereo/multi-view matching or depth sensing devices such as laser scanner, Time-of-Fight (TOF) sensor and the recently launched Microsoft Kinect. The usefulness of the former method relies heavily on how the phenomena such as occlusion, edges, color correlation and so on, are modeled. In certain circumstances, they are able to produce high accuracy depth map with high resolution and wide distance range. However, in texture-less regions, the performances of stereo matching techniques are somewhat limited. Nevertheless, reliable depth maps are usually generated offline and different degrees of human intervention are involved depending on the algorithms being used. On the other hand, most of the depth sensing devices can easily handle the texture-less regions which are contrary to the stereo matching. Unfortunately, existing depth sensing devices suffer from many limitations. For example, conventional laser scanners are too slow for real time usage, and TOF sensors and Kinect are usually poorly calibrated, noisy and limited in resolution. Moreover, their abilities in dealing with transparency materials, object boundaries and wide distance range are not very satisfactory.

From the above discussion, we can see that the depth maps obtained from the stereo/multi-view matching and depth sensing device are indeed complementary to each other. This motives us to develop a new depth map estimation system and approach, which is able to combine the advantages of stereo matching and depth sensing device in order to obtain depth maps with high resolution and accuracy, and yet using much less computational time. The proposed system consists of a high-definition (HD) 3D stereo camera and a Kinect depth sensor. To fully utilize the information obtained from these two different devices, we first calibrate the system using a co-planarity based method. Then, we explore the complementary characteristics of the 3D stereo camera and Kinect, and propose a new method to solve the resultant joint fusion problem. In particular, we derive a fusion framework and derive the probability distribution functions to describe the characteristic of these multimodal depth sensing devices. Moreover, we incorporate into the problem a pixel-wise weighting function which reflects the reliabilities of the stereo camera and Kinect depth sensor. By so doing, a more accurate depth map can be obtained. This reliability fusion concept is first introduced by [12]. However, in their paper, the fusion is completed by TOF and stereo cameras. Therefore, the weighting function is different from us. In the final step, we employ a color image guided depth matting process and 2D polynomial regression (LPR) techniques to further refine the estimated depth map. Simulation results show that the proposed approach is able to offer satisfactory depth maps which significantly outperform their counterparts obtained by either stereo matching or depth sensing device alone.

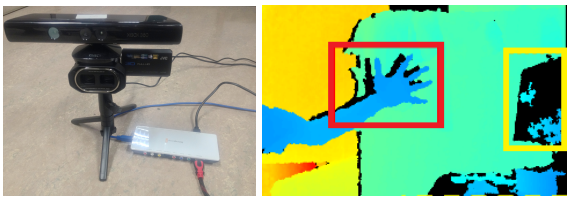
The paper is organized as follows: The setup of the proposed depth estimation system and its calibration procedure are summarized in Section II. Section III is devoted to the joint stereo and Kinect depth fusion algorithm. Experimental results, evaluation and comparison are presented in Section IV. Finally conclusions are drawn in Section V.

II. SYSTEM SETUP AND CALIBRATION

In this section, we introduce the setup of our high resolution depth map estimation system and summarize the methods for calibrating the devices.

A. System Setup

The high resolution depth estimation system constructed, which is shown in Figure 1(a), consists of a Microsoft Kinect and a JVC GS-TD1B FHD 3D Everio camcorder. The Kinect is equipped with an RGB camera and a depth sensor consisting of an infrared camera and an infrared projector. The major features of the Kinect are summarized as follows: (a) It is able to support a distance range from 0.4m to 4 m with an official SDK and further from 0.4m to 8 m with a third party SDK, and (b) it



(a) (b)

Figure 1: (a) The joint stereo and Kinect system for depth Map estimation. (b) Illustration of data missing regions in Kinect's depth map. The red and yellow rectangles are referring the type I and type II missing data.

provides a depth map with 640 x 480 resolution at 30 frames per second (FPS). For the JVC GS-TD1B 3D camera, it provides stereo side-by-side FHD videos in 30 FPS, and it is connected to a Blackmagic-design Intensity shuttle [15] which transfers in real-time the stereo data to PC via HDMI cable for further processing and fusion of the Kinect depth map.

Although the depth maps of Kinect are less noisy than TOF camera [1], there are still many holes and noises which should be suppressed. Apart from the limited sensing range of the Kinect, these holes and noises mainly come from two different sources: I) occlusions between the infrared camera and the infrared projector of the depth sensor and II) material absorption and surface normal direction of objects, as illustrated in Figure 1(b). Moreover, the low resolution of the depth map (640 x 480) will restrict its usage in high resolution applications. In Section III, we will illustrate how to compensate for the above drawbacks of the Kinect using high-resolution stereo cameras.

B. System Calibration

To combine the two different data sources from the Kinect and JVC 3D camcorder, calibration between these devices is required. In the proposed approach, we first calibrate the stereo cameras of the 3D camcorder using standard checkerboard-based method [2], and then calibrate the 3D camcorder and Kinect using co-planarity based method [5]. The basic idea of the latter is to exploit the co-planar property of the calibration board with the help of the JVC 3D camcorder. More precisely, the calibration procedure can be divided into two steps. First, the feature point based method [14] is employed to obtain the initial guess of the intrinsic and extrinsic parameters of the Kinect depth sensor. Then, based on the co-planar property, the system parameters of these two devices can be obtained by solving a non-linear minimization problem. Next, we will consider how to fuse the information offered by the two devices.

III. Joint Stereo and Kinect for High Resolution Depth Estimaion

Recent stereo algorithms mostly employ Markov Random Field (MRF) [7] to model the observation and estimate the stereo correspondences by maximizing the a posterior probability (MAP). One key feature of this MAP-MRF approach is that it provides a systematic framework to integrate the information from multiple sensors. Graph Cuts (GC) and Belief Propagation (BP) are two prevalent methods for approximating the inference in MRF. They are widely used because of their good performances and relatively fast computational time. In what follows, we shall illustrate how the extra Kinect depth data can be incorporated into the MAP-MRF framework and how to solve the resultant fusion problem using a multiscale BP approach.

A. Stereo and Kinect Fusion Problem

The stereo and Kinect fusion problem for depth estimation can be formulated the following MAP problem:

$$P(X \mid Y,Z) \propto \prod_i f_d(x_i,y_i) f_k(x_i,z_i) \prod_{j \in N(i)} f_s(x_i,x_j), \quad (1)$$

where $X = \{x_i, \forall i\}$ denotes the hidden variables associated with the disparities of all pixels, $Y = \{y_i, \forall i\}$ and $Z = \{z_i, \forall i\}$ are observed variables corresponding to the color-based matching cost at specific disparity, and depth value returned by the Kinect respectively, $N(i)$ represents the neighbors of pixel i, denoted by p_i. $f_d(x_i,y_i)$, $f_k(x_i,z_i)$ are local evidences based on the initial pixel-wise matching cost and the measurement from the Kinect depth sensor. $f_s(x_i,x_j)$ represents the smoothness or prior term which incurs discontinuity cost of assigning different disparities x_i and x_j to two neighboring pixels. By taking the logarithm of Eq. (1), it can be seen that the MAP problem is equivalent to minimizing the following objective function

$$E = \sum_i D(x_i,y_i,z_i) + \sum_{j \in N(i)} V(x_i,x_j), \quad (2)$$

where $D(x_i,y_i,z_i) = -\log f_d(x_i,y_i) - \log f_k(x_i,z_i)$ is called the data term and $V(x_i,x_j) = -\log f_s(x_i,x_j)$ is called the smoothness term. In order to produce more accurate fusion results, we propose to use a weighted data term as follows

$$D(x_i,y_i,z_i) = -w_i^d \log f_d(x_i,y_i) - w_i^k \log f_k(x_i,z_i), \quad (3)$$

where w_i^d and w_i^k are the pixel-wise weighting factors for stereo and Kinect depth sensor. They are related to the reliability of each estimated depth pixel resulting from stereo matching and Kinect depth map, which are denoted by H_i^d and H_i^k, respectively. In the proposed fusion framework, we compute these weights factors based on the pixel-wise reliability of both stereo and Kinect as follows:

$$w_i^d = H_i^d / (H_i^d + H_i^k) \text{ and } w_i^k = H_i^k / (H_i^d + H_i^k). \quad (4)$$

Here H_i^d can be computed similar to [12]

$$H_i^d = \begin{cases} 1 - m_i^{1st} / m_i^{2nd} & m_i^{2nd} > T_c \\ 0 & otherwise \end{cases} \quad H_d(p) \in [0,1]. \quad (5)$$

H_i^d quantifies how distinctive the best and the second best matching costs (denoted by m_i^{1st} and m_i^{2nd} respectively) of each p_i. T_c is a small positive threshold to avoid division by zeros. Instead of using absolute difference (AD) as the matching cost, m_i^{1st} and m_i^{2nd} is computed by Birchfield and Tomasi's pixel dissimilarity. This measure of dissimilarity reduces the problem of sampling with little additional

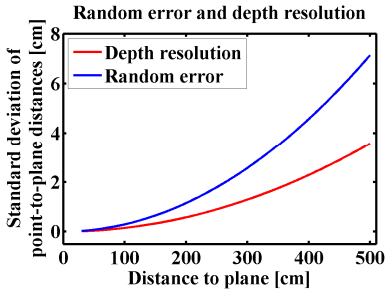


Figure 2: Standard deviations of plane fitting residuals at different distances of the plane to the sensor: theoretical random error σ (red) and depth resolution ζ (blue) [4].

computational complexity, compared to AD. In addition, locally adaptive support-weight approach [8] is employed to overcome matching ambiguities caused by occlusion boundaries, sensor noise and insufficient (repetitive) texture.

The reliability of Kinect depth sensor H_i^k is derived from standard deviations of random error (σ) and depth resolution (ζ) of plane fitting residuals at different plane to sensor distances. Figure 2 shows the calculated σ and ζ plotted against the distance from the plane to the sensor. It can be seen that the errors increase quadratically from 0.5 m distance to the maximum range of the sensor. Therefore, σ and ζ can be used to model the H_i^k as:

$$H_i^k = \begin{cases} 1/\sigma\zeta & \text{if } f_k \neq 0 \\ 0 & otherwise \end{cases} \quad H_i^k \in [0,1]. \tag{6}$$

B. MAP-MRF Configuration and Multiscale BP

From Eq. (1), our MAP-MRF includes three terms, which are the data term from stereo matching, the data term from Kinect depth sensor and the smoothness term. Here we adopt the truncated linear model so that it can be fitted into the efficient multiscale BP framework [10]. We now give the details of these terms. **1)** Data term from stereo matching $f_d(x_i, y_i)$: as we defined in Eq. (1) y_i is the observed variable corresponding to the color-based matching cost m_i at specific disparity. Therefore, the data term for stereo matching $f_d(x_i, y_i)$ can be modeled as:

$$f_d(x_i, y_i) = \min(|x_i - y_i|, \mu_d). \tag{7}$$

Eq. (7) is a truncated linear model, where the cost increases linearly based on the distance between x_i and y_i up to some level. μ_d is the upper bound and it is set to 2.5 in our experiment. **2)** Data term from Kinect depth sensor $f_k(x_i, z_i)$: It encodes the depth consistency between the stereo and the Kinect depth sensor. After joint calibration and rectification, the angles and distances between Kinect and stereo are adjusted. Therefore, the outputs are depth map and stereo images that are row-aligned and rectified. However, the output depth map from Kinect depth sensor is the distance map from sensor to the object surface (plane to sensor distance). In order to corporate with stereo matching, the plane to sensor distance should be transfer to stereo disparity value as:

$$z_{x_i} = lt / z_i + (c_{left} - c_k), \tag{8}$$

where z_{x_i} is the disparity calculated by Kinect depth sensor and x_i denotes the disparity value of stereo matching. z_i is the plane to sensor distance as defined in Eq. (1). l and t are the focal length and baseline between Kinect and the left camera of the stereo pair. c_{left} and c_k are principle points of the left camera of the stereo pair and Kinect depth sensor, respectively. By assuming that the disparity value obtained from Kinect depth sensor should be the same as the stereo matching, we define the cost f_k of the Kinect term as the difference between these two disparities

$$f_k(x_i, z_i) = \min(|x_i - z_{x_i}|, \mu_k), \tag{9}$$

where μ_k is set as the same as μ_d in our experiment. **3)** Smoothness term $f_s(x_i, x_j)$: It is designed based on the magnitude of the difference between x_i and x_j ($j \in N(i)$). We also use a truncated linear model to describe this term

$$f_s(x_i, x_j) = \min(|x_i - x_j|, \mu_s), \tag{10}$$

where μ_s is set to half of the maximum disparity value.

After f_d, f_k and f_s are defined, we employ a multiscale belief propagation (BP) algorithm [10] to solve the MAP-MRF problem. Compare to conventional BP algorithms, this method uses hierarchical technique to obtain a good approximation of the optimal solution with a small fixed number of message passing iterations. In addition, by using Eq. (10), the complexity of the inference can be reduced to linear rather than quadratic in the number of possible labels for each pixel. Interested reader are referred to [10] for details.

C. Depth Map Refinement

We note that the texture of a given object in a neighborhood is usually highly correlated. This is also true for their depth values because they usually arise from the same neighborhood of a physical object. Consequently, the confidence of a given depth pixel is closely related to the correspondent color pixels in the neighborhood. To this end, a two-step approach is considered below to further refine the estimated depth map.

1) Color image guided depth matting process: To explore the connection between color and depth pixels so as to achieve a better visual quality, we shall extend the conventional matting technique of color images to depth images. More precisely, a color image guided depth matting process is proposed to further refine the quality of depth edges under the framework of Bayesian matting [13].

Given the observed color image C and depth image d, the joint color and depth matting problem is to find the proper matting parameters: foreground $F = [F_c \mid F_d]$, background $B = [B_c \mid B_d]$ and opaque α, where F_c and F_d (B_c and B_d) are respectively the foreground (background) in the color image and the correspondent ones in the depth map. Under the framework of Bayesian matting for color images, we have

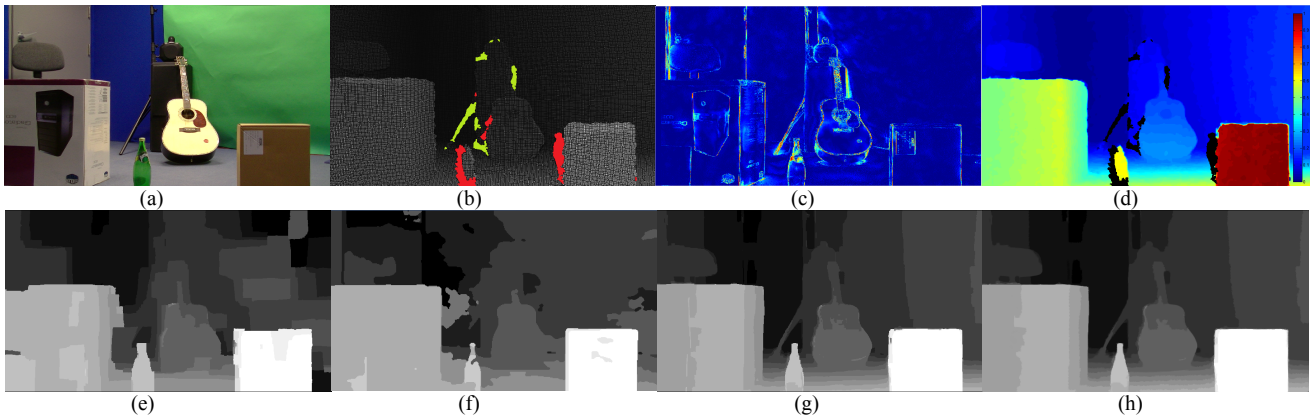


Figure 3: Depth estimation results: (a) reference image; (b) raw depth from Kinect (type I and II invalid regions are highlighted by red and green) (c) H^d map; (d) H^k map; (e) raw matching result from multiscale BP stereo [11] only; (f) matching result from the non-Local Filter (with $\sigma = 0.20$ in [10]); (g) joint stereo and Kinect fusion result and (h) joint stereo and Kinect fusion result with depth map refinement.

$$\begin{aligned} &\arg\max_{F,B,\alpha} P(F,B,\alpha \mid C,d) \\ &= \arg\max_{F,B,\alpha} P(C,d \mid F,B,\alpha)P(F)P(B)P(\alpha)/\big(P(C)P(d)\big), \\ &= \arg\max_{F,B,\alpha} L(C,d \mid F,B,\alpha) + L(F) + L(B) + L(\alpha) \end{aligned} \quad (11)$$

where $L(.) = \log P(.)$ is the log likelihood. Note that the log likelihood terms are modeled similar to the Bayesian matting approach [13], and $P(C)$ and $P(d)$ are dropped because they are constant with respect to the optimization parameters. The maximization problem in Eq. 11 can be divided into two sub problems to iteratively solve F, B and α, which is similar to the maximization problem described in Bayesian matting [13]. After the matting parameters $\{F_d, B_d, \alpha\}$ is obtained, the depth map can be refined and the edges in the depth map can be matted to reflect our confidence on the actual depth values.

2) 2D LPR smoothing: To further reduce possible image noise arising from low texture, occlusion, etc, the depth maps should be further smoothed. Here, we adopt 2D LPR with adaptive bandwidth selection [9] for smoothing the estimated depth map after matting. It is particularly useful in preserving the discontinuity at object boundaries while performing smoothing at flat areas.

IV. Experimental Results

We now present and evaluate the experimental results of the proposed system and algorithm. More precisely, the quality of the depth estimation from stereo matching [10, 11], the Kinect depth sensor and our joint stereo and Kinect fusion are compared using an indoor complex scene. The left view of the JVC 3D camera was set as the reference view and the resolution is 720p (HD). In this paper, 5 message passing iterations per level and 5 levels in total was used in the multiscale BP framework. The processing time was approximately 0.5 second on an Intel i7 920 CPU-based computer with 4GB RAM and GTX295 GPU acceleration.

The image from the reference view is shown in Figure 3(a). It contains texture-less regions, transparency objects and type I and II factors which will cause holes in depth map captured by Kinect, as shown in Figure 3(b). Since the resolution of the reference image is much higher than the Kinect depth sensor, the depth data cannot cover every pixel of the reference image. Moreover, depth values of object boundaries obtained by Kinect are unstable and it will cause significant artifacts in IBR and other applications. Finally, a lot of type I and type II invalid regions (holes) can be found from the raw depth map of Kinect.

We can see from Figures 3(c) and (d) that the proposed reliability maps can effectively capture the strengths of the two devices and demonstrate their complementary nature. H^d is computed based on the distinctiveness between the first and second matching cost of each pixel. Heavy textured regions will have higher H^d in the fusion and low texture regions such as the wall and texture-less object surfaces will trend to depend on the Kinect results. H^k is obtained based on Figure 2, which reflects the reliability of the depth information according to distance from the object to the Kinect depth sensor.

Figure 3(e) shows the depth map from multiscale BP stereo [11] only. Unlike Kinect result in Figure 3(b), the depth of the green bottle is successfully estimated in this stereo matching method. However, it is erroneous in texture-less regions and there are large ambiguities of assigning depth values to pixels around object boundaries. Figure 3(f) shows a matching result obtained from newly proposed method which is based on non-local cost aggregation (non-Local Filter) and non-local disparity refinement method [10]. We can see that the green bottle can be observed and the depth discontinuities are well preserved. Obviously, the result of [10] outperforms that of [11] but both methods fail to reconstruct thin structure such as the guitar bar and tripod in the scene. In addition, we found that both stereo methods cannot successfully reconstruct the back of the chair. Figure 3(g) shows our fusion result, which is of high quality and resolution as compared to those of multiscale BP stereo only, non-Local filter approach and Kinect's result. The texture-less regions are well handled and holes of Figure 3(b) are also filled by reasonable values. However the boundaries of objects are not sharp enough and there are still some noise in the depth map. Therefore, the two-step depth map refinement is employed to further improve the depth map and the result is shown in Figure

3(h). Compare to Figure 3(g), the object boundaries in Figure 3(h) are better preserved and the noises are efficiently suppressed.

V. CONCLUSION

A new high resolution depth estimation system using joint stereo vision and Kinect has been presented. Methods for calibrating the devices and joint depth estimation using the MRF-MAP framework are presented. The problem is solved using the multiscale BP framework and further processed by joint color and depth matting and depth map filtering using 2D LPR. Experimental results show that our system outperforms either the conventional stereo vision or Kinect alone.

REFERENCES

[1] J. Smisek, M. Jancosek and T. Pajdla, "3D with Kinect," in *IEEE Workshop on Consumer Depth Cameras for Computer Vision,* 2011.

[2] Z. Y. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 22, no. 11, pp. 1330-1334, Nov. 2000.

[3] H. Y. Shum, S. C. Chan and S. B. Kang, *Image-based rendering,* NY: Springer-Verlag, 2007.

[4] K. Khoshelham and S. Oude Elberink, "Accuracy and resolution of Kinect depth data for indoor mapping applications," *Sensors,* vol. 12, no. 2, pp. 1437-1454, 2012.

[5] D. Herrera, C., J. Kannala and J. Heikkila, "Joint depth and color camera calibration with distortion correction," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 99, no. PrePrints, May. 2012.

[6] C. Zhang and Z. Zhang, "Calibration between depth and color sensors for commodity depth cameras," in *Int. Workshop Hot Topics in 3D, in conjunction with ICME*, 2011.

[7] L. Zhang and S. Seitz, "Parameter estimation for MRF stereo," in *Proc. IEEE Comput. Soc. Conf. CVPR,* vol. 2, Aug. 2005, pp. 288-295.

[8] K. J. Yoon and I. S. Kweon, "Locally adaptive support-weight approach for visual correspondence search," in *Proc. IEEE Intl Conf. Computer Vision and Pattern Recognition*, pp. 924-931, 2005.

[9] Z. Y. Zhu, S. Zhang, S. C. Chan and H. Y. Shum, "Object-based rendering and 3D reconstruction using a moveable image-based system," *IEEE Trans. Circuits Syst. Video Technol*., vol. 22, no. 10, pp.1405-1419, Oct. 2012.

[10] Q. X. Yang, "A non-local cost aggregation method for stereo matching," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Aug. 2012, pp. 1402-1409.

[11] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *Intl. J. comput. vision*, vol. 70, no. 1, pp.41-54, 2006

[12] J. J. Zhu, L. Wang, R. G. Yang, J. E. Davis and Z. G. Pan, "Reliability fusion of time-of-fight depth and stereo geometry for high quality depth maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, July 2011.

[13] Y. Chuang, B. Curless, D. H. Salesin and R. Szeliski, "A Bayesian Approach to Digital Matting," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Dec. 2001, vol. II, pp. 264-271.

[14] [Online]. Available: http://nicolas.burrus.name/.

[15] [Online]. Abailable: http://www.blackmagicdesign.com/.